

Stat 240

Dr. Dave Campbell
dac5@SFU.ca

What is data?

information.

Note that quality can vary considerably (see courses in experimental design and survey sampling)

What is science?

<https://www.merriam-webster.com/dictionary/science>

“knowledge or a system of knowledge covering general truths or the operation of general laws especially as obtained and tested through the scientific method and concerned with the physical world and its phenomena.”

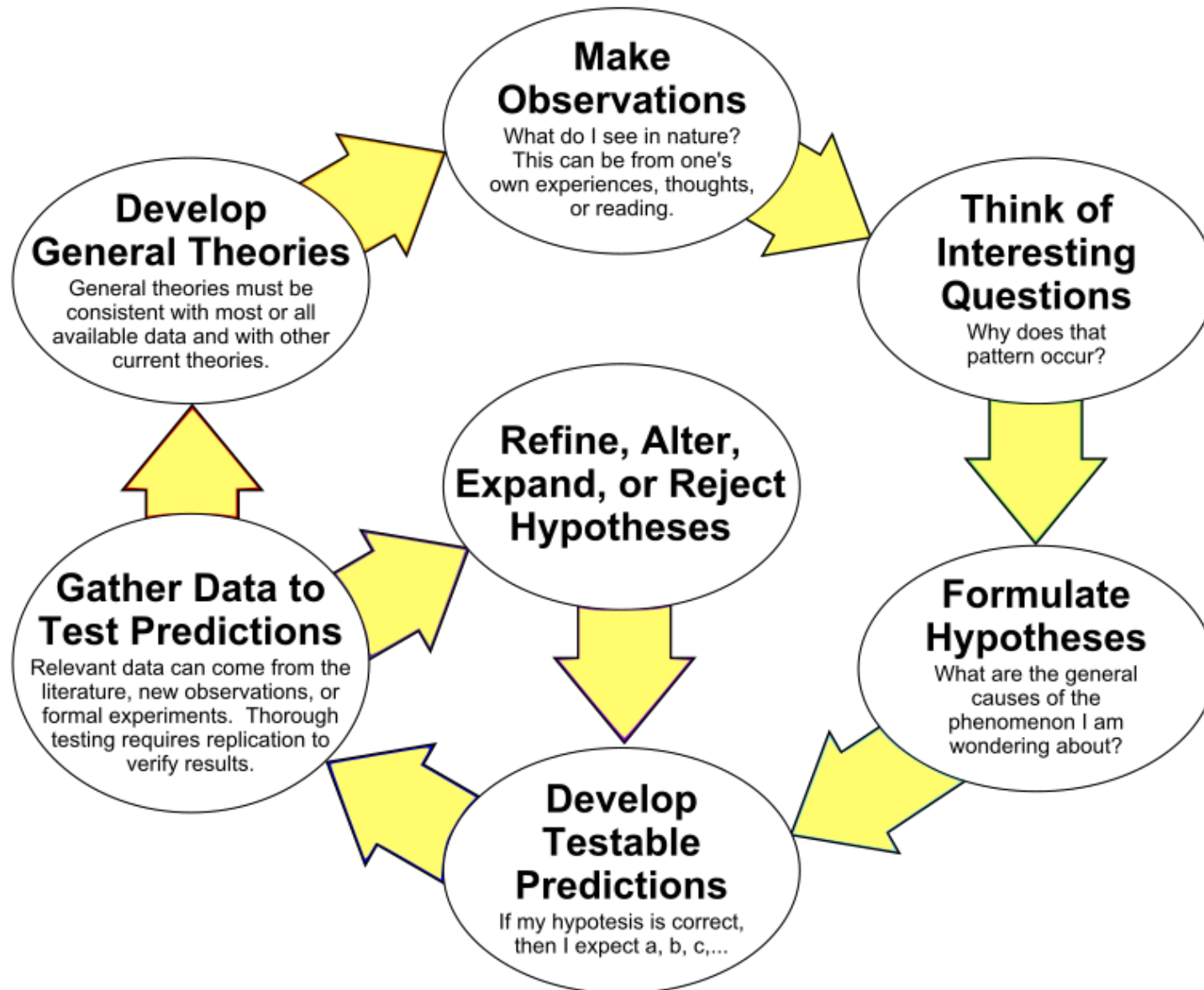
What is the scientific method?

https://en.wikipedia.org/wiki/Scientific_method

“A body of techniques for investigating phenomena, acquiring new knowledge, or correcting and integrating previous knowledge.”

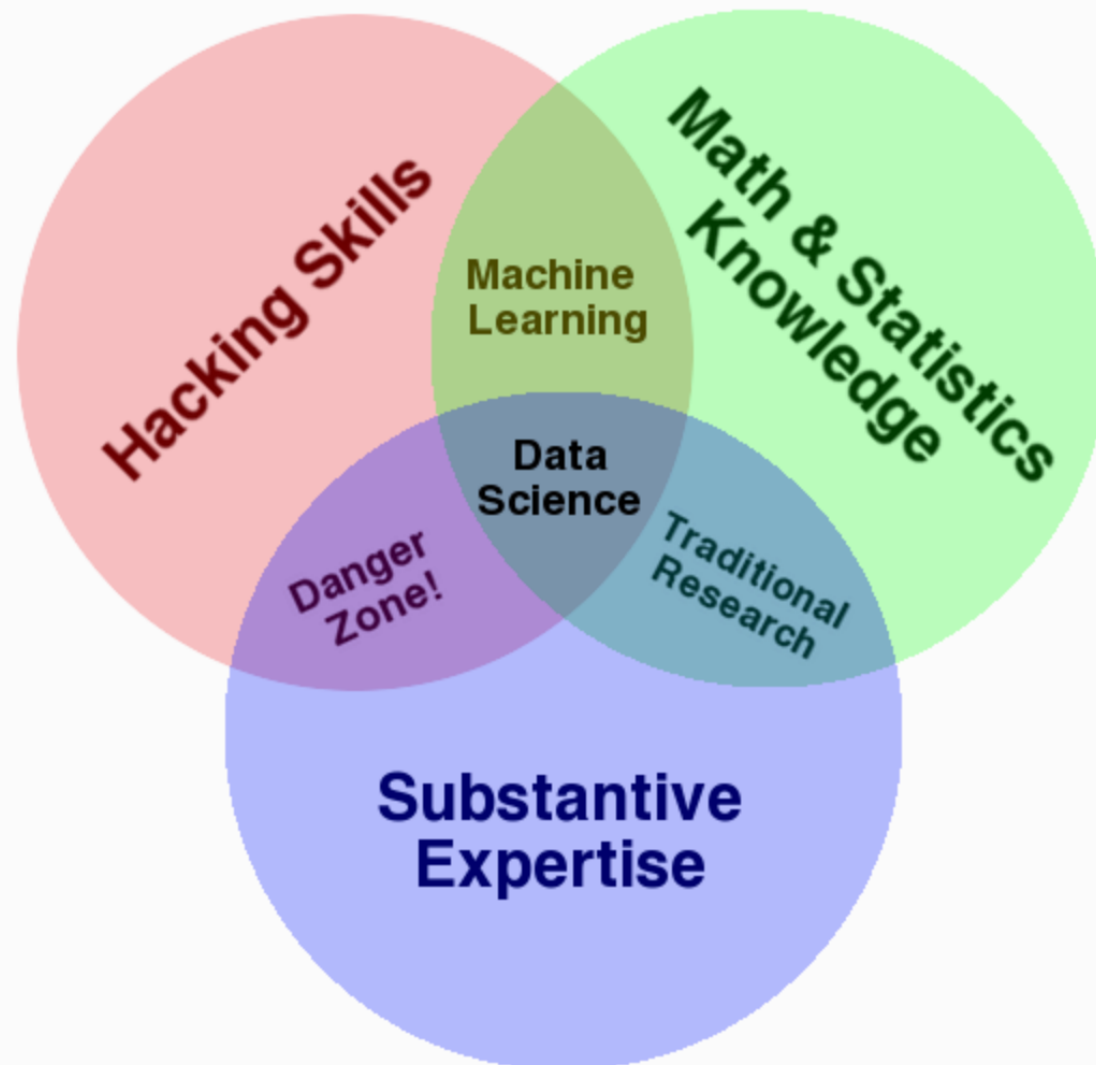
https://en.wikipedia.org/wiki/Scientific_method

The Scientific Method as an Ongoing Process



Drew Conway - The Data Science Venn Diagram
<http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>

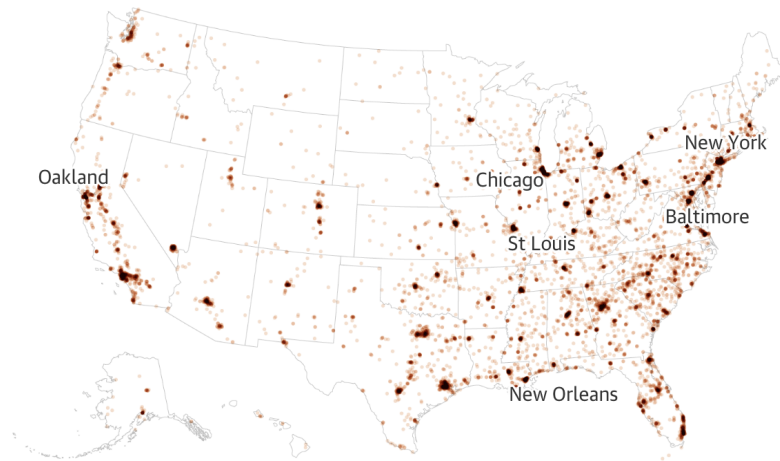
What Is Data Science?



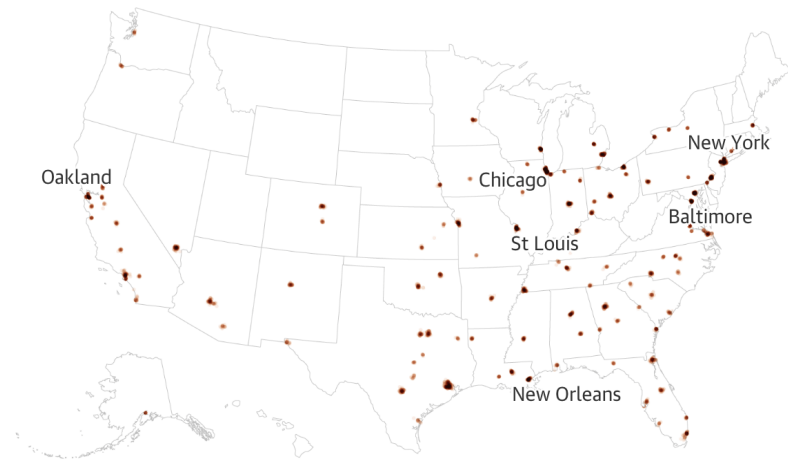
<https://www.theguardian.com/us-news/ng-interactive/2017/jan/09/special-report-fixing-gun-violence-in-america>

Data Science is about pulling together data to answer questions, provide insights, or help with evidence based decisions.

In 2015, there were more than 13,000 gun homicides throughout the US ...



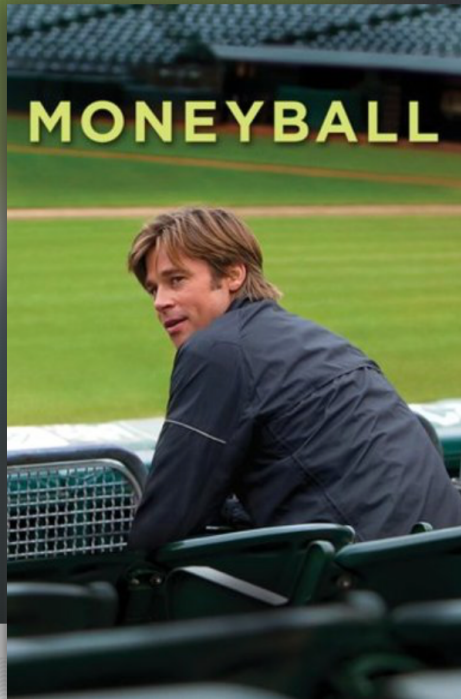
... but half of those deaths were in just 127 cities, which contain almost a quarter of the population



<http://www.sonypictures.com/movies/moneyball/>

Moneyball

NOW AVAILABLE



Trailer

[WATCH NOW](#)



Disc & Digital
Purchase
Options
[VIEW ALL](#)

ABOUT MONEYBALL

Oakland A's general manager Billy Beane (Brad Pitt) challenges the system and defies conventional wisdom when he is forced to rebuild his small-market team on a limited budget. Despite opposition from the old guard, the media, fans and their own field manager (Philip Seymour Hoffman), Beane - with the help of a young, number-crunching, Yale-educated economist (Jonah Hill) - develops a roster of misfits...and along the way, forever changes the way the game is played.

How did Michael Burry (Christian Bale) win?

He looked at the data.

The Big Short (2015) ★ 7.8_{/10} 222,360 ☆ Rate This

R | 2h 10min | Biography, Comedy, Drama | 23 December 2015 (USA)



2:05 | Trailer 39 VIDEOS | 127 IMAGES

Four denizens in the world of high-finance predict the credit and housing bubble collapse of the mid-2000s, and decide to take on the big banks for their greed and lack of foresight.

Director: [Adam McKay](#)

Writers: [Charles Randolph](#) (screenplay), [Adam McKay](#) (screenplay) | [1 more credit](#) »

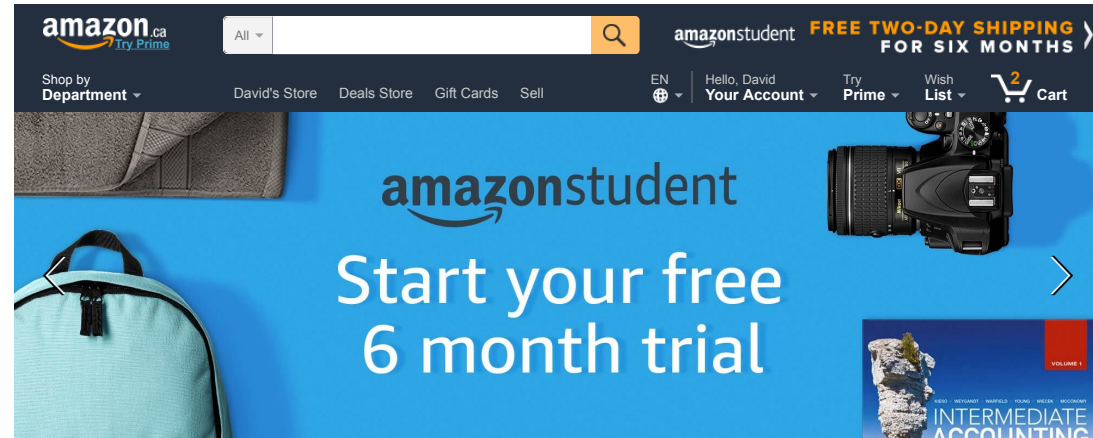
Stars: [Christian Bale](#), [Steve Carell](#), [Ryan Gosling](#) | [See full cast & crew](#) »

81 Metascore From metacritic.com	Reviews 399 user 435 critic	Popularity 136 (▲ 18)
---	---	---------------------------------

Won 1 Oscar. Another 35 wins & 80 nominations. [See more awards](#) »

Data Science Successes

Recommendation systems



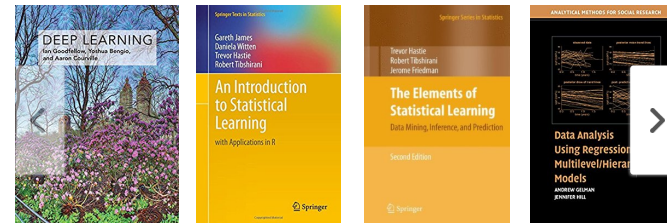
FF Foo Fighters Fanzone
Sponsored · 🌐

Do You Love Dave Grohl!
Comment "Yes, I need this" if you want one!
If you love it, get it here: <http://tiny.cc/Foo-Fighter6>

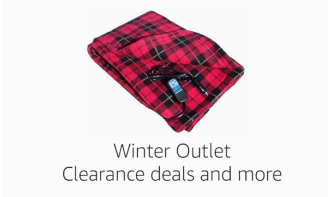
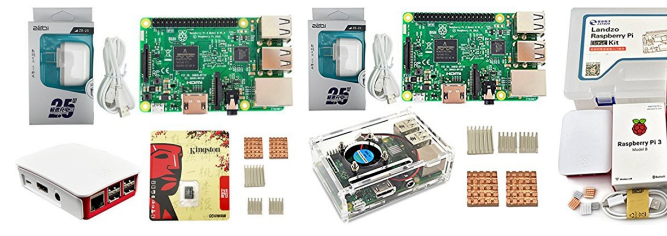


👍❤️👹 12.1K 1.5K Comments 1.3K Shares

Related to items you've viewed [See more](#)



More items to consider [See more](#)

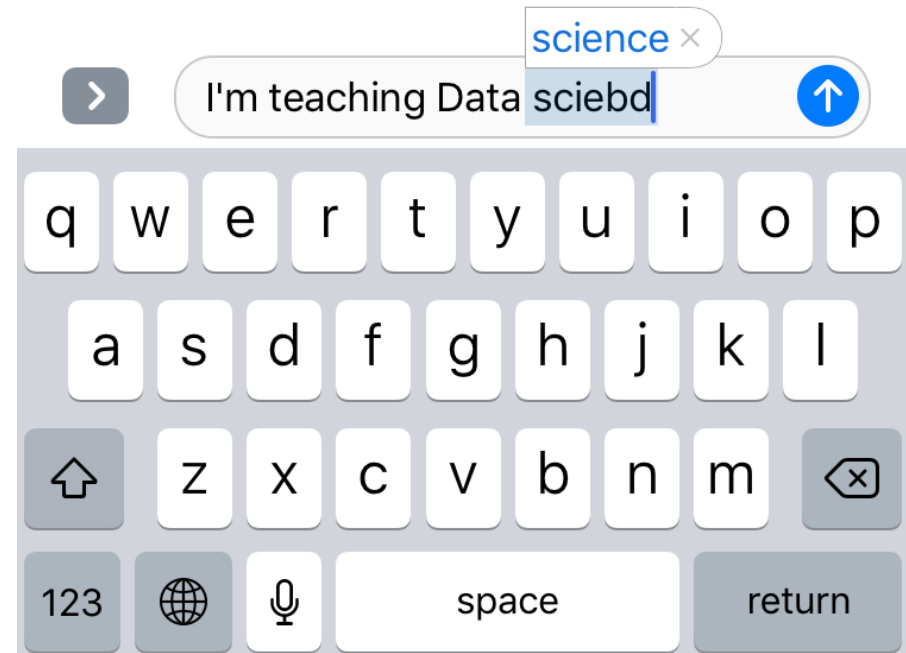


Inspired by your shopping trends



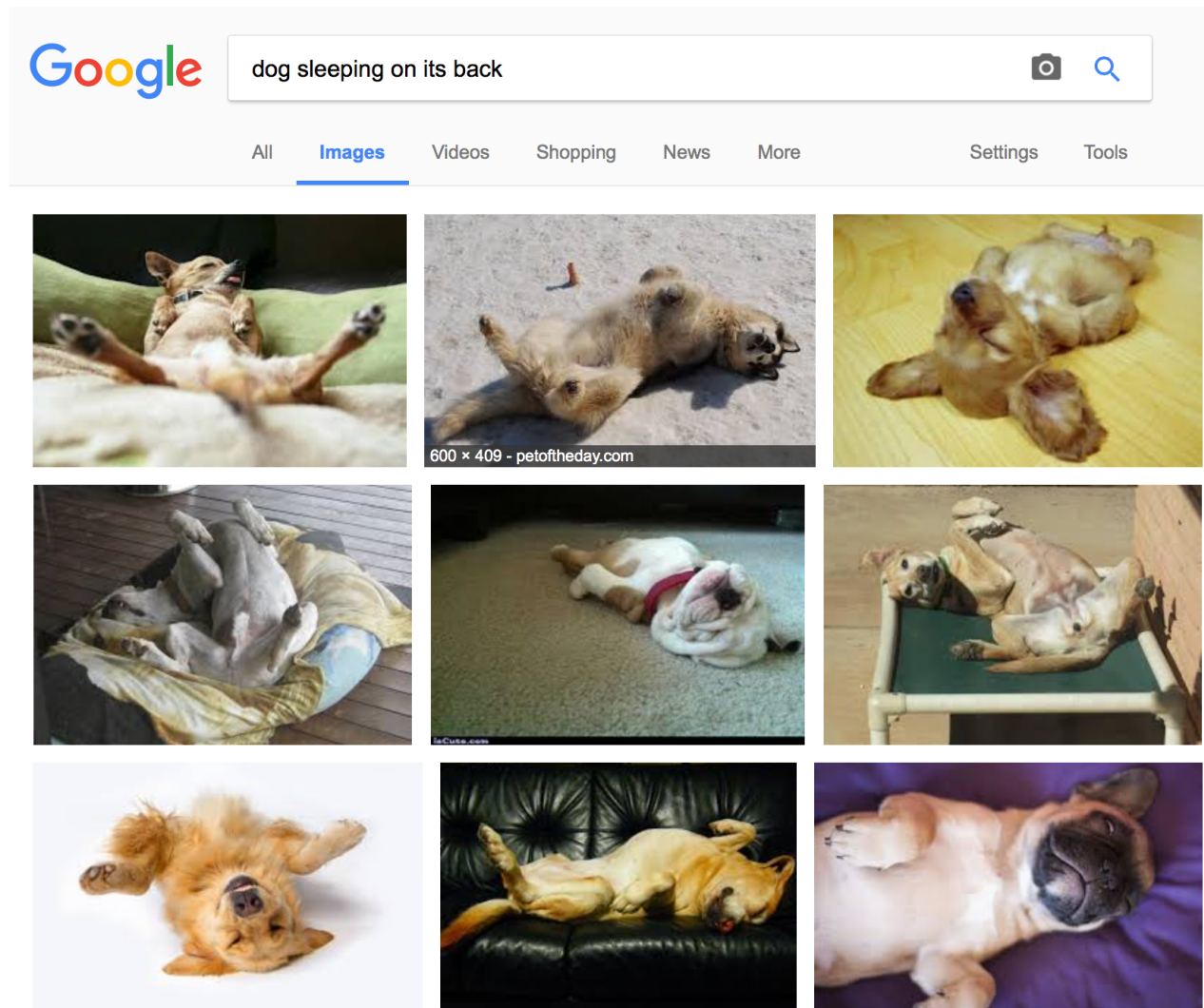
Data Science Successes

autocorrect

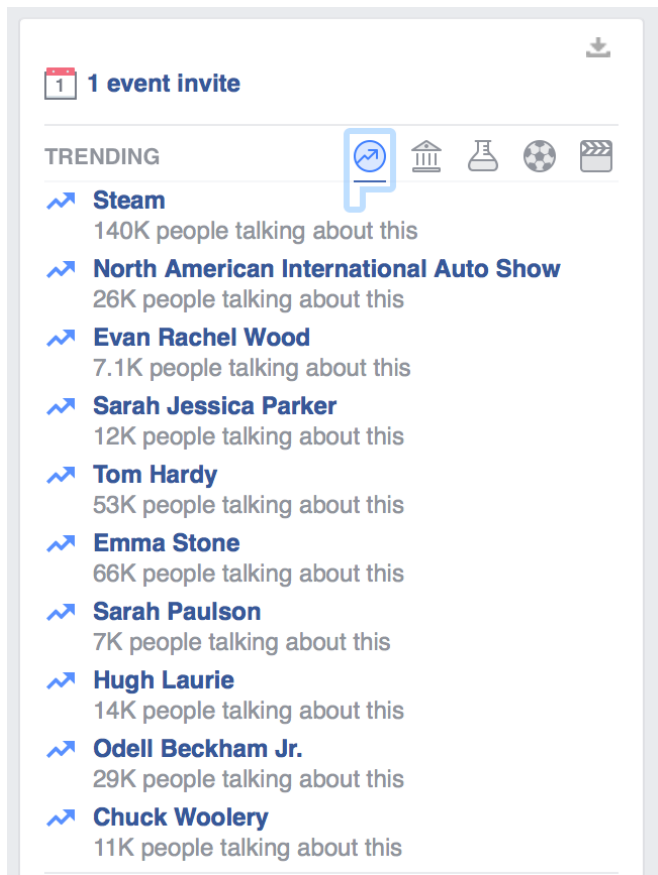


Data Science Successes

Search
Engines



Describing what's happening now

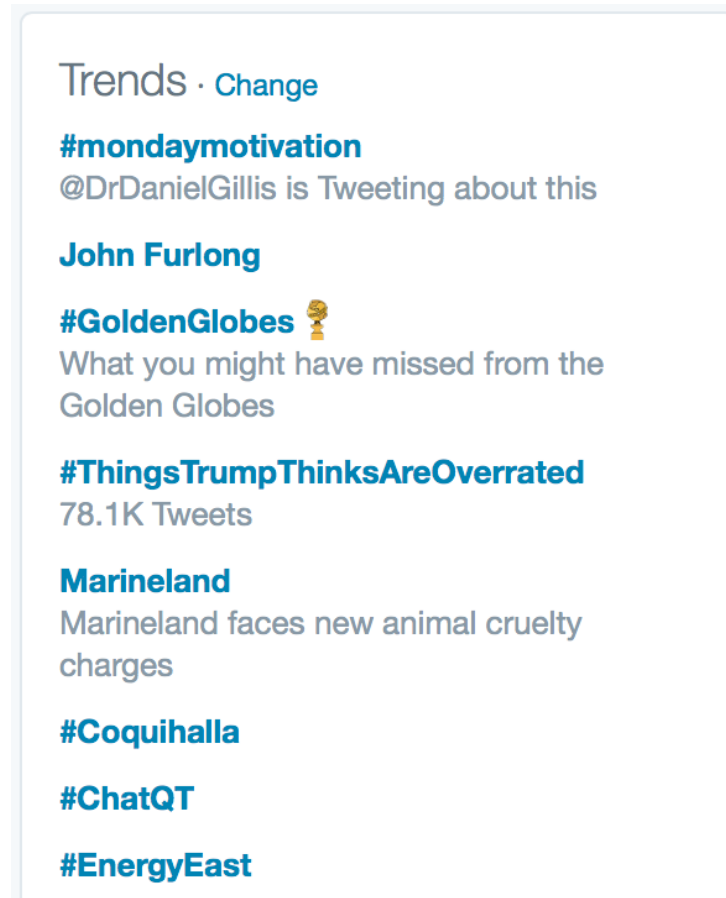


1 event invite

TRENDING

- Steam**
140K people talking about this
- North American International Auto Show**
26K people talking about this
- Evan Rachel Wood**
7.1K people talking about this
- Sarah Jessica Parker**
12K people talking about this
- Tom Hardy**
53K people talking about this
- Emma Stone**
66K people talking about this
- Sarah Paulson**
7K people talking about this
- Hugh Laurie**
14K people talking about this
- Odell Beckham Jr.**
29K people talking about this
- Chuck Woolery**
11K people talking about this

The screenshot shows a social media trending page. At the top, there is a notification for '1 event invite'. Below that is a 'TRENDING' section with a list of items. The first item, 'Steam', is highlighted with a blue speech bubble icon. To the right of the trending list are several icons representing different categories: a building, a flask, a soccer ball, and a film strip.



Trends · Change

- #mondaymotivation**
@DrDanielGillis is Tweeting about this
- John Furlong**
- #GoldenGlobes** 🏆
What you might have missed from the Golden Globes
- #ThingsTrumpThinksAreOverrated**
78.1K Tweets
- Marineland**
Marineland faces new animal cruelty charges
- #Coquihalla**
- #ChatQT**
- #EnergyEast**

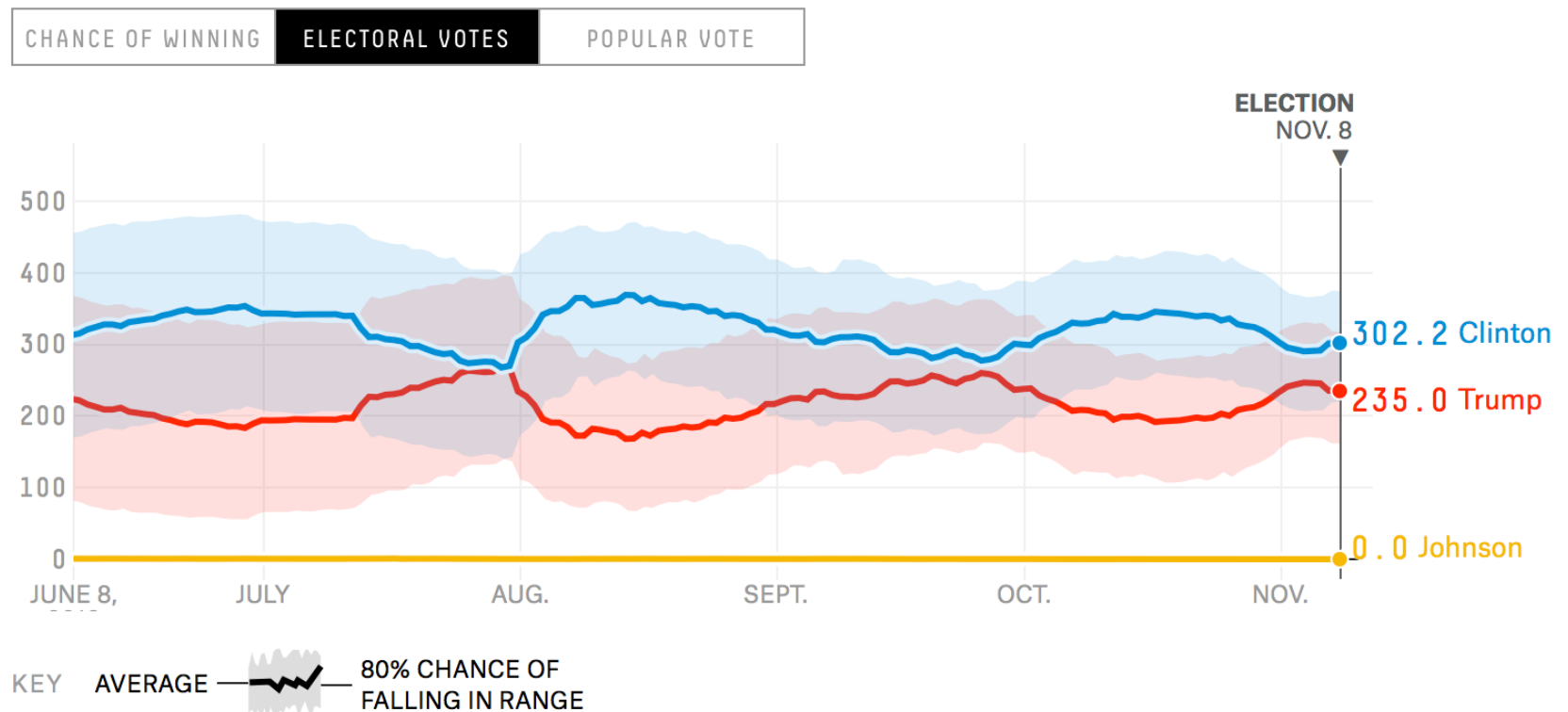
The screenshot shows a social media trends page. At the top, it says 'Trends · Change'. Below that is a list of trending topics. The first item is '#mondaymotivation' with a note that '@DrDanielGillis is Tweeting about this'. The second item is 'John Furlong'. The third item is '#GoldenGlobes' with a trophy icon and a note that 'What you might have missed from the Golden Globes'. The fourth item is '#ThingsTrumpThinksAreOverrated' with '78.1K Tweets'. The fifth item is 'Marineland' with a note that 'Marineland faces new animal cruelty charges'. The sixth item is '#Coquihalla'. The seventh item is '#ChatQT'. The eighth item is '#EnergyEast'.

Prediction

https://projects.fivethirtyeight.com/2016-election-forecast/?ex_cid=rrpromo

How the forecast has changed

We'll be updating our forecasts every time new data is available, every day through Nov. 8.



Other successes

Fraud detection

Self driving cars

Airline route planning (predicting flight delays, route popularity, customer preferences...)

More case studies

January 21-22, 2017 SFU, Surrey Campus.

<https://www.healthhackathon.ca>

Data Science

Data Science involves the full process of obtaining, warehousing, and cleaning data

Data Science involves the full process of seeking value in data, asking domain specific questions, summarizing and analyzing data, and producing evidence based decisions.

This course focuses on data acquisition and cleaning with some work on data summaries and analysis.

Commonalities in a sample of 28 jobs in Data Science in Canada

More jobs for those with a graduate degree right now. Junior positions are opening up everywhere.

R, Python, SQL, Hadoop, Java, Tableau,

Data Visualization, Exploratory analysis (Stat 341/342)

Advanced Statistical models, regression, GLM, experimental design, hypothesis testing (stat 350, 430, 410, 475...)

Machine Learning methods, random forests, neural nets, clustering, classification, text analytics, optimization (Stat 440, Stat 452)

Data acquisition, wrangling and exploratory data analysis. Data warehousing, often high volume (Stat 240)

Strong communication skills, presentation skills, sharing insights and technology transfer (Stat 300)

Software Engineering skills, cloud computing, dashboarding, parallel / map reduce

marketing, management, understanding customers, formulate problem statements, project management for senior roles

Often coming up with your own hypotheses, then experimenting and testing to develop new insights to communicate to the team

Implement algorithms , optimization, and analysis

Data Science

Data Scientist is the “Sexiest Job of the 21st Century” - Harvard Business Review

<https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century>

Data Science

Data scientists as “a job title that barely existed three years ago but since has become one of the hottest corners of the high-tech labor market”

<http://www.wsj.com/articles/academic-researchers-find-lucrative-work-as-big-data-scientists-1407543088>

The October 2015 report "Closing Canada's Big Data Talent Gap" by Canada's Big Data Consortium (made up of the Government of Canada, SFU, and other academic and industrial leaders) states "Bold promises have been made for Big Data and Analytics (i.e., Data Science): exceptional customer insights; better decision-making; improved productivity and performance; and product and service innovation.... But the promise of Big Data and Analytics faces a key constraint: a talent gap that is felt across all of Canada's regions, sectors, and industries." [4] The report further outlines that "Canada's Big Data Talent Gap is estimated between 10,500 and 19,000 professionals with deep data and analytical skills, such as those required for roles like Chief Data Officer, Data Scientist, and Data Solutions Architect."

http://www.ryerson.ca/content/dam/provost/pdfs/ryerson_ccbdtg.pdf

In this course

Part 1: Intro to ouR software (wks 1 & 2)

Part 2: Using R to deal with large databases
(wks 3 - 5)

Part 3: Obtaining text data from an API (wks 6 &
7)

Part 4: Scraping data from webpages (wks 9-11)

Lecture vs lab

lecture; listen, ask questions, learn concepts, see some stuff in action.

lab: try things, do stuff, start assignments for the week, get directed assistance.

In this course

Part 1: Intro to our software (wks 1 & 2)

basic commands, how to do basic stuff

Part 2: Using R to deal with large databases (wks 3 - 5)

accessing a remote server, obtaining the pieces of the dataset that you need, some data cleaning

Part 3: Obtaining text data from an API (wks 6 & 7)

making queries for real time data, dealing with text data

Part 4: Scraping data from webpages (wks 9-11)

dealing with text and numbers, html, extracting useful pieces of data, automation

Grading

Participation 4%

Weekly lab assignment and their pre-lab submission (9-10): 50%

Midterm (Mar 6, i.e. wk8) 16% ← This will be (at least partially) in the computer lab

Final 30% ← after the midterm I will decide if this will be on computer during the lab on Apr 3 or a “regular” final held during the usual exam period

Office hours & Contacting Dr. Dave Campbell

K10564 Thursdays 11-12.

Other times by appointment: dac5@SFU.ca

We'll use canvas soon

About me

Dr. David Campbell

Science background: BSc in Environmental Science

Statistical modelling: MSc

Mathematical and Statistical Theory: PhD.

Ongoing: Co-organize a reading group on Data Science

Meet with / collaborate with industry

Research is mainly Statistical Methodology

Today

Intro to R and Rstudio

On lab machines use Rstudio

At home 1st install R: <https://www.r-project.org>

Then install Rstudio <https://www.rstudio.com>

Rstudio is a user interface for R

R vs RStudio

Deriving insights from data

How did students do in my last class?

Data:

Final grades from my Stat 285 last term
(sorted lowest to highest)

[8,] 0.5411765
[9,] 0.5411765
[10,] 0.5529412
[11,] 0.5529412
[12,] 0.5647059
[13,] 0.5764706
[14,] 0.6235294
[15,] 0.6352941
[16,] 0.6470588
[17,] 0.6470588
[18,] 0.6823529
[19,] 0.6823529
[20,] 0.6941176
[21,] 0.6941176
[22,] 0.7058824
[23,] 0.7176471
[24,] 0.7294118
[25,] 0.7294118
[26,] 0.7294118
[27,] 0.7294118
[28,] 0.7411765
[29,] 0.7529412
[30,] 0.7647059
[31,] 0.7647059
[32,] 0.7764706
[33,] 0.7764706
[34,] 0.7764706
[35,] 0.7882353
[36,] 0.7882353
[37,] 0.7882353
[38,] 0.7882353
[39,] 0.8117647
[40,] 0.8117647
[41,] 0.8117647
[42,] 0.8235294
[43,] 0.8352941
[44,] 0.8352941
[45,] 0.8588235
[46,] 0.8705882
[47,] 0.8705882

Commands

“=” vs “<-“

a vector: `c(1, 2, 3, 4, 5)`

finding elements `a[2]`

all but the second element `a[-2]`

logicals, ‘<’, ‘>’, ‘==’, ‘max’, ‘min’

`hist(x,nbins, xlim, main, xlab)`

`#comments`